CSE 312
Foundations of Computing II

Lecture 22: Pagerank, Tail Bounds and Other Loose Ends

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Anna R. Karlin

Slide Credit: Based on Stefano Tessaro's slides for 312 19au
incorporating ideas from Ryan O'Donnell, Alex Tsun, Rachel Lin, Hunter Schafer & myself ☺

PageRank: Some History

The year was 1997

- Bill Clinton in the White House
- Deep Blue beat world chess champion (Kasparov)

The internet was not like it was today. Finding stuff was hard!

- In Nov 1997, only one of the top 4 search engines actually found itself when you searched for it

The Problem

Search engines worked by matching words in your queries to documents.

Not bad in theory, but in practice there are lots of documents that match a query.

- Search for Bill Clinton, top result is ‘Bill Clinton Joke of the Day’
- Susceptible to spammers and advertisers

The Fix: Ranking Results

- Start by doing filtering to relevant documents (with decent textual match).
- Then **rank** the results based on some measure of ‘quality’ or ‘authority’.

Key question: How to define ‘quality’ or ‘authority’?

Enter two groups:

- Jon Kleinberg (professor at Cornell)
- Larry Page and Sergey Brin (Ph.D. students at Stanford)

Both groups had the same brilliant idea

Larry Page and Sergey Brin (Ph.D. students at Stanford)

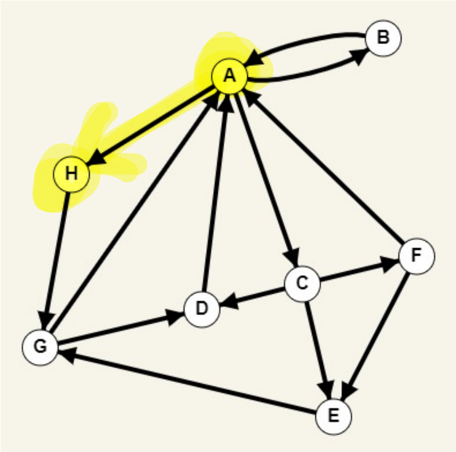
- Took the idea and founded Google, making billions

Jon Kleinberg (professor at Cornell)

- MacArthur genius prize, Nevanlinna Prize and many other academic honors

PageRank - Idea

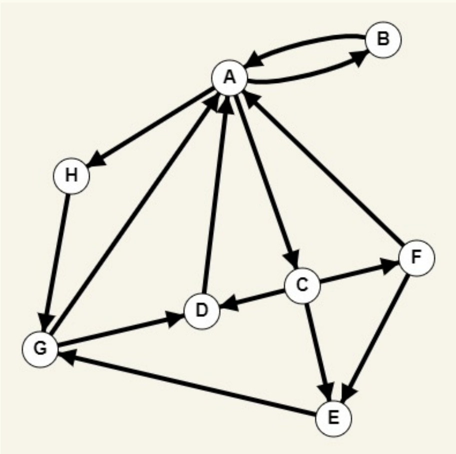
Take into account directed graph structure of the web. Use **hyperlink analysis** to compute what pages are high quality or have high authority. Trust the internet itself define what is useful via its links.



PageRank - Idea

Idea 1: think of each link as a citation “vote of quality”

Rank pages by in-degree?



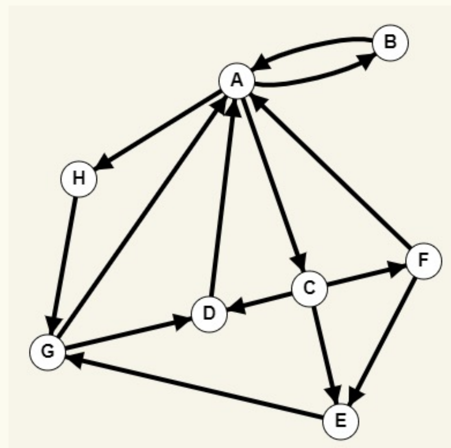
PageRank - Idea

Idea 1: think of each link as a citation "vote of quality"

Rank pages by in-degree?

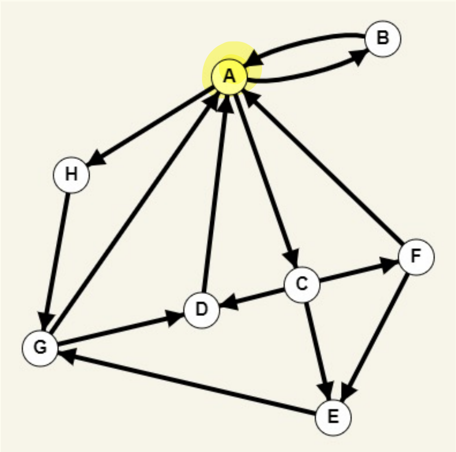
Problems:

- Spamming
- Some linkers not discriminating
- Not all links created equal



PageRank - Idea

Idea 2: perhaps we should weight the links somehow and then use the weights of the in-links to rank pages

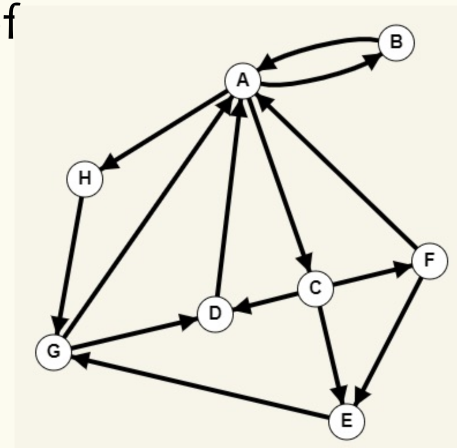


Inching towards Pagerank

Web page has high quality if it's linked to by lots of high quality pages.

A page is high quality if it links to lots of high quality pages

recursive definition!



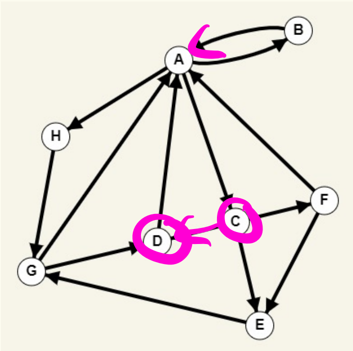
Inching towards Pagerank

- If web page x has d outgoing links, one of which goes to y, this contributes 1/d to the importance of y.
- But we want to take into account the importance of x.

q_x: quality of page x.

$q_A = q_B \cdot 1 + q_F \frac{1}{2} + q_D \cdot 1 + q_G \frac{1}{2}.$

$q_D = q_C \cdot \frac{1}{3} + q_G \cdot \frac{1}{2}$



$p_{ij} = \begin{cases} \frac{1}{\text{outdegree}(i)} & \text{if } i \to j \text{ is hyperlink} \end{cases}$

$(q_1, \ldots, q_n)$

Gives the following equations

Idea: Use the transition matrix defined by a random walk on the web $\boldsymbol{P}$ to compute quality of webpages. Namely, find $\boldsymbol{q}$ such that

$$qP = q$$

Look familiar?

This is the stationary distribution for the Markov chain defined by a random surfer. Starts at some node (webpage) and randomly follows a link to another.

– Use stationary distribution of her surfing patterns after a long time as notion of quality

Issues with PageRank

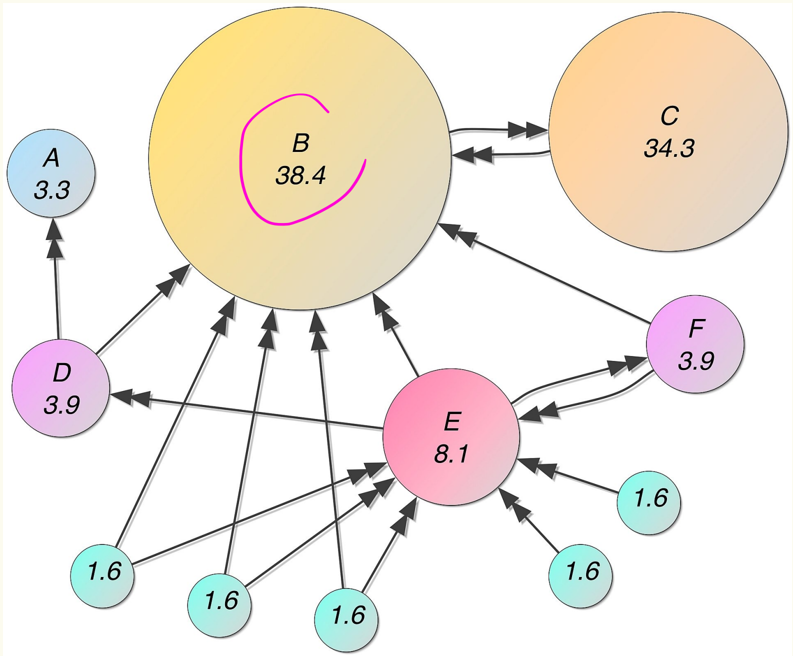
- How to handle dangling nodes (dead ends)?
- How to handle Rank sinks – group of pages that only link to each other?

Both solutions can be solved by "teleportation"

Final PageRank Algorithm

- Make a Markov Chain with one state for each webpage on the internet with the transition probabilities $\boldsymbol{P}_{ij} = \frac{1}{outdeg(i)}$.
- Use a modified random walk. At each point in time, if the surfer is at some webpage x.
  - With probability p take a step to one of the neighbors of x (equally likely)
  - With probability $1 - p$, "teleport" to a uniformly random page in the whole internet.
- Compute stationary distribution $\boldsymbol{\pi}$ of this perturbed Markov chain.
- Define the PageRank of a webpage i as the stationary probability π_i.
- Find all pages with decent textual match to search and then order those pages by PageRank!

PageRank - Example

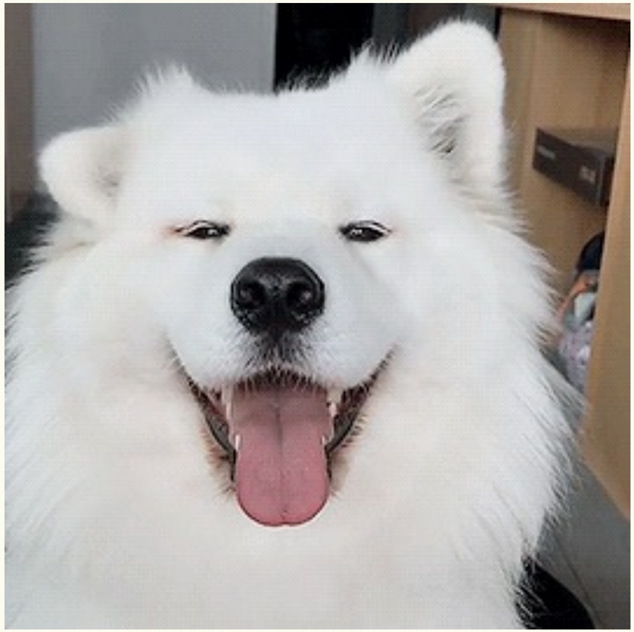


It Gets More Complicated

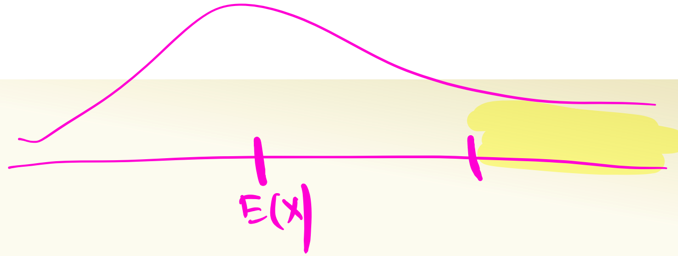
While this basic algorithm was the defining idea that launched Google on their path to success, this is far from the end to optimizing search.

Nowadays, Google has a LOT more secret sauce to ranking pages most of which they don't reveal for 1) competitive advantage and 2) avoid gaming their algorithm.

Brain Break



Tail Bounds (Idea)



Bounding the probability a random variable is far from its mean. Usually statements of the form:

$$\Pr(X \geq a) \leq b$$
$$\Pr(|X - E[X]| \geq a) \leq b$$

Useful tool when

- An approximation that is easy to compute is sufficient
- The process is too complex to analyze exactly

A gambling game

- With probability 0.99, you pay me \$10
- With probability 0.01, I pay you \$1000

- Do you want to play?

$E(\text{you gain}) = 1000 \cdot 0.01 - 10 \cdot 0.99 = 10¢$

$E(\text{my gain}) = -10¢$

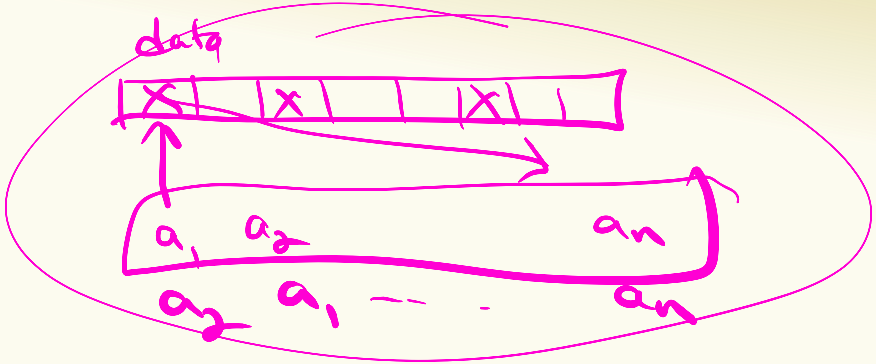
$\Pr(\text{you gain} \geq \text{exp}) = 0.01 \rightarrow 10^{-100}$

$\Pr(\text{my gain} \geq \text{exp}) = 0.99$

Takeaway

- A random variable might almost never be at least its expectation.
- Similarly, a random variable might almost always be at least its expectation.

Changes to minimum



Compute-Min

min := ∞

For t := 1 to [redacted]

If data [t] < min

min := data [t]

print (“The new minimum is “, min) *

Suppose that the data array contains n distinct numbers.

All permutations are equally likely

E(number of times line * is executed) = $1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n} = H_n$

$\approx \ln(n)$

$E(X) = \sum_{k=1}^{n} k \Pr(X=k) \geq \sum_{k=n/2}^{n} k \Pr(X=k)$

tries get a new min

$\geq \frac{n}{2} \sum_{k=n/2}^{n} \Pr(X=k)$

$\Pr(X \geq \frac{n}{2})$

- E(X) about ln(n)
- Possible that Pr $(X \geq n)$ = 0.99?
- Possible that Pr $(X \geq n/2)$ = 0.99?

$E(X) \geq 0.99 \frac{n}{2} \gg \ln n$

$E(X) = E(X \mid X \geq \frac{n}{2}) \Pr(X \geq \frac{n}{2}) + E(X \mid X < \frac{n}{2}) \Pr(X < \frac{n}{2})$

$\ln n$ $\quad \geq 0$

$\geq E(X \mid X \geq \frac{n}{2}) \Pr(X \geq \frac{n}{2})$

$$\Pr\left(X \geq \frac{n}{2}\right) \leq \frac{2\ln n}{n}$$

Agenda

- **Markov's Inequality**
- Chebyshev's Inequality
- The Law of Large Numbers

$$E(X) = \sum_{x \in \Omega_X} x \cdot \Pr(X = x)$$

$$\Pr(X \geq 10\,E(X)) \leq \frac{E(X)}{10\,E(X)}$$

$(t = 10\,E(X))$

Markov's Inequality

Theorem. Let X be a random variable taking only non-negative values. Then, for any $t > 0$,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}.$$

$$\mathbb{P}\big(X \geq t \cdot \mathbb{E}(X)\big) \leq \frac{1}{t}.$$

Incredibly simplistic – only requires that the random variable is non-negative and only needs you to know <u>expectation</u>. You don't need to know **anything else** about the distribution of X.

Markov's Inequality – Proof

Theorem. Let X be a (discrete) random variable taking only non-negative values. Then, for any $t > 0$,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}.$$

$$\mathbb{E}(X) = \sum_{x} x \cdot \mathbb{P}(X = x)$$

$$= \sum_{x \geq t} x \cdot \mathbb{P}(X = x) + \sum_{x < t} x \cdot \mathbb{P}(X = x)$$

≥ 0

$$\geq \sum_{x \geq t} x \Pr(X = x)$$

$W + 100$

Markov's Inequality – Proof

Theorem. Let X be a (discrete) random variable taking only non-negative values. Then, for any $t > 0$,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}.$$

$$\mathbb{E}(X) = \sum_{x} x \cdot \mathbb{P}(X = x)$$

$$= \sum_{x \geq t} x \cdot \mathbb{P}(X = x) + \sum_{x < t} x \cdot \mathbb{P}(X = x)$$

≥ 0 because $x \geq 0$ whenever $\mathbb{P}(X = x) \geq 0$ (takes only non-negative values)

$$\geq \sum_{x \geq t} x \cdot \mathbb{P}(X = x)$$

$$\geq \sum_{x \geq t} t \cdot \mathbb{P}(X = x) = t \cdot \mathbb{P}(X \geq t)$$

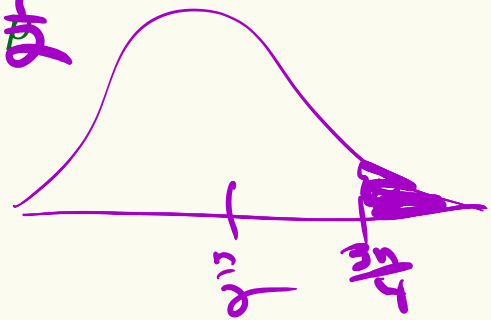
Follows by re-arranging terms …

Example – Binomial Random Variable

Markov's inequality
$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}$.

Let X be Binomial RV with parameters. n, p

$$\mathbb{E}(X) = \frac{n}{2}$$



What is the probability that $X \geq \frac{3n}{4}$?

Markov's inequality: $\mathbb{P}\left(X \geq \frac{3n}{4}\right) \leq \frac{4}{3n} \cdot \frac{n}{2} = \frac{2}{3}$

Can we do better?

Agenda

- Markov's Inequality
- Chebyshev's Inequality
- The Law of Large Numbers

Using variance

- If we know more about the random variable, e.g. its variance, we can get a better bound!

Chebyshev's Inequality

Markov's inequality
$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}$.

Theorem. Let X be a random variable. Then, for any $t > 0$,

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}.$$

Proof: Define $Z = X - \mathbb{E}(X)$

Definition of Variance

$$\mathbb{P}(|Z| \geq t) = \mathbb{P}(Z^2 \geq t^2) \leq \frac{\mathbb{E}(Z^2)}{t^2} = \frac{\mathrm{Var}(X)}{t^2}$$

$|Z| \geq t$ iff $Z^2 \geq t^2$

Markov's inequality ($Z^2 \geq 0$)

Example – Binomial Random Variable

Chebychev's Inequality
$\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}$.

Let X be Binomial RV with parameters. $n, p = 0.5$

$\mathbb{E}(X) = \frac{n}{2}$ $\quad\quad$ $Var(X) =$

What is the probability that $X \geq \frac{3n}{4}$ *?*

Chebychev's inequality: $\mathbb{P}\left(X \geq \frac{3n}{4}\right) \leq$

Markov's inequality: $\mathbb{P}\left(X \geq \frac{3n}{4}\right) \leq \frac{4}{3n} \cdot \frac{n}{2} = \frac{2}{3}$

Chevychev gives us a pretty good bound. But still not great.

Chebychev's inequality: $\mathbb{P}\left(X \geq \frac{3n}{4}\right) \leq \frac{4}{n}$

For 1000 flips, probability of getting at least 750 heads is at most 0.02

The truth for n = 1000:

$$\mathbb{P}\left(X \geq \frac{3n}{4}\right) < 0.000000000000000000000000000000000000000000067$$

The Law of Large Numbers

Chebychev's Inequality
$\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}$.

(Weak version) Let $X_1, X_2, \dots, X_n$ be i.i.d. random variables with mean μ, and let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

$$\lim_{n\to\infty} P\left(|\bar{X} - \mu| > \epsilon\right) = 0.$$

Tail Bounds

Useful for approximations of complex systems. How good the approximation is depends on the actual distribution and the context you are using it in.

- Usually loose upper-bounds are okay when designing for worst-case

Generally, the more you know about your random variable the better tail bounds you can get.